

Response under 37 C.F.R. 1.116

Applicant: Maria Castellanos et al.

Serial No.: 09/944,919

Filed: August 31, 2001

Docket No.: 10007912-1

Title: METHOD AND SYSTEM FOR MINING A DOCUMENT CONTAINING DIRTY TEXT

REMARKS

The following remarks are in response to the Final Office Action mailed March 4, 2004. With this Response, claims 1-12 and 15-35 remain pending in the application and are presented for reconsideration and allowance.

Claim Rejections under 35 U.S.C. § 103

The Examiner rejected claims 1-12 and 15-35 under 35 U.S.C. § 103(a) as being unpatentable over Domini et al. U.S. Patent No. 6,085,206 (Domini) in view of Zamora U.S. Patent No. 4,965,763 (Zamora). Applicant submits that Domini, either alone or in view of Zamora, fails to teach or suggest the invention of independent claims 1, 11, 21, and 31.

Independent claim 1 recites a computer-implemented method for mining a document containing dirty text. The method includes removing an instance of dirty text within the document to produce a cleaned document having a content. The method also includes performing a data mining operation on said cleaned document thereby deriving relevant information from said cleaned document and providing a summary of the content of said document.

Domini is directed to removing/correcting dirty text in a document, including correction of both spelling and grammatical construction in a document at the same time. Domini specifically defines dirty text as that text which has not been spell checked and/or that has not been grammar checked (See Domini column 9, lines 43-48). Furthermore, Domini describes that after a sentence has been grammar checked, it is marked as clean text (column 9, lines 49-53).

Zamora merely recites a computer method for automatic extraction of commonly specified information from business correspondence. The method utilizes a parametric information extraction (PIE) system to identify fields of a business document such as frame slots for a business correspondence or list of business correspondence closing phrases (See Zamora, Fig. 3 and Fig. 5). The identified fields disclosed are limited to "standardized forms" (Col. 3, l. 36) such as "author, date, recipient, address, subject statement . . ." (Col. 3, ll. 23-24).

Domini fails to disclose **performing a data mining operation on a cleaned document**. The Examiner incorrectly concluded that correction of grammar is a form of data mining. Similar to Applicant, Domini defines the extraction of grammatical constructions as removal of dirty text. In further contrast, **Domini fails to disclose performing a data mining operation on the clean document thereby deriving relevant information from said clean document and providing a summary of the content of said document**. Again, Domini is directed to correcting spelling and grammar in a document and does not disclose performing a data mining operation that results in providing a summary of the content of the document.

Further, Zamora fails to disclose removing an instance of dirty text within said document to produce a cleaned document having a content. Further, Domini fails to disclose performing a data mining operation on said cleaned document thereby derive a relevant from said cleaned document and providing a summary of the content of said document. In contrast, Zamora uses a parametric information extraction system to identify fields of a business document, such as author, dates, recipient, address, etc. Since neither Domini nor Zamora teach or suggest performing a data mining operation on a cleaned document thereby deriving relevant information from said cleaned document and providing a summary of the content of said document, one skilled in the art could not apply the teachings of Domini in view of Zamora and arrive at the present invention of independent claim 1. In fact, Zamora teaches away from cleaning a document prior to performing a data mining operation, since Zamora is triggering on specific and expected business letter fields like closing phrases and headers.

Accordingly, Applicant respectfully requests that the above rejection under 35 U.S.C. § 103(a) should be withdrawn. Dependant claims 2-10 depend directly or indirectly upon independent claim 1. Accordingly, dependant claims 2-10 are also allowable over the art of record.

Claim 11 recites a computer system. The computer system includes a bus, a memory unit coupled to the bus, and a processor coupled to the bus. The processor executes a method for mining a document containing dirty text. The method includes producing a cleaned document having a content including performing a general cleaning of the document by removing an instance of dirty text within the document including instances of misspelling and

grammatical errors, performing a domain and task-specific cleaning of the document including removing instances of computer code and tables to produce a cleaned document. A data mining operation is performed on the cleaned document including providing a summary of the content of the document.

For similar reasons as stated above with reference to independent claim 1, Applicant believes independent claim 11 to be allowable over Domini either alone or in view of Zamora. Further, neither Domini nor Zamora teach or suggest a two-step cleaning process. Specifically, neither Domini nor Zamora teaches or suggests performing a general cleaning of said document by removing an instance of dirty text within said document including instances of misspelling and grammatical errors, followed by performing a domain and task specific cleaning of said document including removing instances of computer codes and tables. As such, one skilled in the art could not combine the teachings of Domini in view of Zamora and arrive at the present invention of independent claim 11.

Dependent claim 12, 15, 19 and 20 depend either directly or indirectly upon independent claim 11. Accordingly, these dependent claims are allowable over the art of record.

Independent claim 21 recites a computer-usable medium performing steps including mining a document containing dirty text. The method includes removing an instance of dirty text within the document to produce a cleaned document having a content. The method also includes performing a data mining operation on said cleaned document thereby deriving relevant information from said cleaned document and providing a summary of the content of said document.

For similar reasons as stated above with reference to independent claim 1, Applicant believes independent claim 21 to be allowable over Domini in view of Zamora. Accordingly, Applicant respectfully requests that the above rejection under 35 U.S.C. § 103(a) should be withdrawn. Dependant claims 22-30 depend either directly or indirectly upon independent claim 21. Accordingly, these dependent claims are allowable over the art of record.

Independent claim 31 recites a computer-implemented method for mining a document containing dirty text. The method includes producing a cleaned document having a content comprising performing a general cleaning of said document by removing one or more instance of dirty text within said document including instances of misspelling and

grammatical errors, and performing a domain and task specific cleaning of said document including removing instances of computer codes and tables. A data mining operation is performed on said cleaned document, including determining a sentence score for each sentence of said cleaned document and ranking the sentences from highest to lowest based on the sentence score. A summary of the content of the document is generated using the highest ranked sentences.

For similar reasons as stated above with reference to independent claims 1, 11 and 21, Applicant believes independent claim 31 to be allowable over Domini in view of Zamora. In addition, nothing in the art of record teaches or suggests determining a sentence score for each sentence of said cleaned document and ranking the sentences from highest to lowest based on the sentence score to provide a summary based on the highest ranked sentences, after completion of a two-step cleaning process. The Examiner references a scoring in Zamora (Col 2, ll. 24-31), but that is limited to determining how many occurrences there are of a user-defined search term in various documents that are being searched and then ranking the various documents. Again, Zamora fails to disclose determining a sentence score for each sentence of said cleaned document and ranking the sentences as claimed by Applicant. One skilled in the art could not combine the teachings of Domini in view of Zamora and arrive at the present invention of independent claim 31.

Accordingly, Applicant respectfully requests that the above rejection under 35 U.S.C. § 103(a) should be withdrawn. Dependant claims 31-35 depend directly or indirectly upon independent claim 31, they are also allowable over the art of record. Further, neither Domini nor Zamora teach or suggest determining a sentence score for each sentence including applying a keyword technique to each sentence (claim 32); applying a location technique to each sentence (claim 33); applying a semantic similarity technique to each sentence (claim 34); wherein the semantic similarity technique comprises generating a vector associated with each sentence, and comparing each vector to every other vector, including defining a co-sign of an angle between two vectors and using the co-sign of the angle between two vectors to determine whether sentences represented by the two vectors are semantically related (claim 35).

Response under 37 C.F.R. 1.116

Applicant: Maria Castellanos et al.

Serial No.: 09/944,919

Filed: August 31, 2001

Docket No.: 10007912-1

Title: METHOD AND SYSTEM FOR MINING A DOCUMENT CONTAINING DIRTY TEXT

Therefore, Applicants respectfully request reconsideration and withdrawal of the 35 U.S.C. § 103(a) rejection to claims 1-12 and 15-35, and request entry of this response and allowance of these claims.

Response under 37 C.F.R. 1.116

Applicant: Maria Castellanos et al.

Serial No.: 09/944,919

Filed: August 31, 2001

Docket No.: 10007912-1

Title: METHOD AND SYSTEM FOR MINING A DOCUMENT CONTAINING DIRTY TEXT

CONCLUSION

In view of the above, Applicant believes independent claims 1, 11, 21 and 31 and the claims depending therefrom are in condition for allowance. Allowance of these claims is respectfully requested.

The Examiner is invited to contact the Applicants representative at the below-listed telephone number to facilitate prosecution of this application.

Any inquiry regarding this Amendment and Response should be directed to either Steven E. Dicke at Telephone No. (612) 573-2002, Facsimile No. (612) 573-2005 or Howard Boyle at Telephone No. (281) 518-9645, Facsimile No. (218) 514-8332. In addition, all correspondence should continue to be directed to the following address:

Hewlett-Packard Company
Intellectual Property Administration
P.O. Box 272400
Fort Collins, Colorado 80527-2400

Respectfully submitted,

Maria Castellanos et al.,

By their attorneys,

DICKE, BILLIG & CZAJA, PLLC
Fifth Street Towers, Suite 2250
100 South Fifth Street
Minneapolis, MN 55402
Telephone: (612) 573-2003
Facsimile: (612) 573-2005

Date: April 27, 2004
SED:jan

Steven E. Dicke
Steven E. Dicke
Reg. No. 38,341

CERTIFICATE UNDER 37 C.F.R. 1.8: The undersigned hereby certifies that this paper or papers, as described herein, are being deposited in the United States Postal Service, as first class mail, in an envelope address to: Mail Stop AF, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on this 27 day of April, 2004.

By Steven E. Dicke
Name: Steven E. Dicke